

Economics 143: Problem Set 1

1. Assume that you are in charge of the central monetary authority in a mythical country. You are given the following historical data on the quantity of money and national income (both in millions of dollars):

Year	Quantity of money	National income	Year	Quantity of money	National income
1987	2.0	5.0	1992	4.0	7.7
1988	2.5	5.5	1993	4.2	8.4
1989	3.2	6.0	1994	4.6	9.0
1990	3.6	7.0	1995	4.8	9.7
1991	3.3	7.2	1996	5.0	10.0

- (a) Plot these points on a scatter diagram. Then estimate the regression of national income Y on the quantity of money X and plot the line on the scatter diagram.
- (b) How do you interpret the intercept and slope of the regression line?
- (c) If you had sole control over the money supply and wished to achieve a level of national income of 12.0 in 1997, at what level would you set the money supply?
2. (a) Assume that least-squares estimates are obtained for the relationship $Y = a + bX$. After the work is completed, it is decided to multiply the units of the X variable by a factor of 10. What will happen to the resulting least-squares slope and intercept?
- (b) Generalize the result of part (a) by evaluating the effects on the regression of changing the units of X and Y in the following manner:

$$Y^* = c_1 + c_2 Y \quad X^* = d_1 + d_2 X$$

What can you conclude?

3. What happens to the least-squares intercept and the slope estimate when all observations on the independent variable are identical? Can you explain intuitively why this occurs?
4. The Capital Asset Pricing Model (CAPM) implies that the mean return of a risky asset, such as a stock, in excess of the return of a risk-free asset, such as a treasury bill, is proportional to the mean excess return of the market portfolio, i.e. the return of a portfolio that contains all risky assets in excess of the return of a risk-free asset. That is:

$$E[R_a - R_f] = \beta E[R_m - R_f] \tag{1}$$

where R_a is the return of the risky asset, R_f is the return of the risk-free asset, and R_m is the return of the market portfolio. The coefficient of proportionality β is known in the literature as the risky asset's *beta*. Let $r_a \equiv R_a - R_f$ and $r_m \equiv R_m - R_f$ be the excess returns on the risky asset and the market portfolio, respectively. We can write equation (1) above as:

$$E[r_a] = \beta E[r_m]$$

In particular, the basic implication of the CAPM model is that the risky asset's beta is equal to:

$$\beta = \frac{Cov(r_a, r_m)}{Var(r_m)} = \frac{E[(r_a - \mu_a)(r_m - \mu_m)]}{E[(r_m - \mu_m)^2]} \tag{2}$$

where $\mu_r = E[r_a]$ and $\mu_m = E[r_m]$. Equation (1) suggests that one may estimate IBM's beta and test whether the CAPM model holds by running the OLS regression:

$$r_{at} = \alpha + \beta r_{mt} + \varepsilon_t$$

and testing if $\alpha = 0$. At the class web site you will find three files that contain monthly returns from January 1978 till December 1987 ($n = 120$ observations in total) for a risky asset, "Ibm" (IBM stock), a risk-free asset, "Rkfree" (30-day US Treasury Bill), and a market portfolio, "Market".

- (a) Construct the series $r_a \equiv R_a - R_f$ and $r_m \equiv R_m - R_f$ and plot them against time in the same graph. Comment on the similarities and differences between the two series.
- (b) Report the OLS estimates of α and β and their standard errors.

Economics 143: Problem Set 2

1. Construct 95 percent confidence intervals for the estimated parameters for Exercise 1.1. Can you reject the null hypothesis that $\beta = 0$? $\beta = 1$?
2. Prove that R^2 for the two-variable regression is unchanged if a linear transformation is made on both variables; that is, $Y^* = a_1 + a_2Y$, $X^* = b_1 + b_2X$.
3. When dealing with time series data it is often observed that economic variables exhibit *time trends*, i.e. a tendency to grow (positive trend) or decline (negative trend) over time. For example, in the table below, the variable Y , the number of deaths of children under age 1 (in thousands) exhibits a negative time trend, while the variable X , the consumption of beer (in bulk barrels) exhibits a positive time trend.

Year	1935	1936	1937	1938	1939	1940	1941	1942	1943	1944	1945	1946
Y	60	62	61	55	53	60	63	53	52	48	49	43
X	23	23	25	25	26	26	29	30	30	32	33	31

The presence of such trends may produce spurious results when trying to estimate the relationship between two or more variables. It is thus common practice to *detrend* the variables first. If a variable appears to grow (or decline) linearly with time, it is reasonable to *fit a linear time trend*. A linear time trend may be fitted to X (or Y) by calculating a LS regression of X (or Y) on time t . The detrended values are then the residuals from that regression.

- (a) Fitting a trend requires choosing an origin and a unit of measurement for the time variable. For example, if the origin is set at mid-1935 and the unit of measurement is 1 year, then the year 1942 corresponds to $t = 7$, and so forth for the other years. If the origin is set at end-1940 (beginning of 1941) and the unit of measurement is 6 months, then 1937 corresponds to $t = -7$. Show that any computed trend value $\hat{Y}_t = \hat{a} + \hat{b}t$ is unaffected by the choice of origin and unit of measurement, where \hat{a} and \hat{b} are the LS estimates. Do \hat{a} and \hat{b} change when we change the origin and unit of measurement?
 - (b) Calculate the correlation coefficient between X and Y and the slope coefficient of the regression of Y on X . Is there a positive or negative linear relationship between the two variables? What is the estimated mean of the number of children deaths if the government prohibited beer consumption? By how much does this number increase or decrease if beer consumption increases by one bulk barrel?
 - (c) Even though we may expect in this example some correlation between X and Y (drunk driving may cause children's deaths), the results from (b) above may be exaggerating the strength of the relationship. To see if this conjecture is correct, calculate the correlation coefficient of the *detrended* values of the two variables.
 - (d) Run the least squares regression of the *detrended* values of Y on the *detrended* values of X and a constant. Calculate the standard errors for the two coefficients (slope and intercept) and carry out a t -test of the hypothesis that the slope coefficient is equal to 0. Interpret your results.
4. The data set EX45.TXT available at the class web site contains the following variables:

RTDR: RETAIL SALES: DURABLE GOODS STORES, 1967:1-1995:8 M, TOTAL, S.A

IVRDR: RETAIL INVENTORIES: DURABLE GOODS 1967:1-1995:7 M, (MIL\$,EOM,SA)

FYCP: INTEREST RATE: COMMERCIAL PAPER, 6-MONTH 1967:1-1995:8 M, (% PER ANNUM,NSA)

LEH: AVG HR EARNINGS OF PROD WKRS: TOTAL 1967:1-1995:8 M, PRIVATE NONAGRIC (\$,SA)

PUCD: CPI-U: DURABLES (82-84=100,SA) 1967:1-1995:8 M

Use this data set to estimate the following model of durable goods sales, and interpret all quantities in the Stata output:

$$SD_t = \beta_1 + \beta_2 DI_{t-6} + \beta_3 IS_{t-1} + \beta_4 I_{t-1} + \beta_5 E_{t-1} + \beta_6 P_{t-1} + \varepsilon_t$$

where

SD = monthly retail sales of durable goods (millions of dollars)

DI = retail inventory of department stores in durable goods (millions of dollars)

IS = Inventory sales ratio for all durable goods

I = open market rate on prime 6-month commercial paper (percent)

E = Average hourly gross earnings of workers (dollars)

P = Consumer Price Index for durable goods (1983=100)

Economics 143: Problem Set 3

1. Using the data set WAGE.TXT (in ASCII format), do the following:
 - (a) Create and save a Stata data set that contains the following variables in that order: WAGE FEMALE NWHITE UNION EDU EXPE AGE.
 - (b) Provide summary statistics for the variables of your data set.
 - (c) Discuss how you would look for evidence on discrimination in wages by gender, race, and union status.
 - (d) Write down a multivariate regression model of wage as a (linear) function of the rest of the variables.
 - (e) Estimate the model by OLS using Stata.
 - (f) Interpret your results.
 - (g) Obtain the vector of predicted dependent variable values and the residuals and plot the latter as a function of the predicted values.
 - (h) Obtain the (estimated) variance-covariance matrix of the OLS estimates.

2. You have the position of political analyst at a local TV station. In view of upcoming congressional elections, you are asked to comment on the results from the 2000 presidential elections. You have data on the percentage of votes received by Democratic candidates among all votes cast for House of Representatives candidates for each one of the 50 states. In addition, for each one of the states you have data on the unemployment rate, and you know whether Al Gore appeared there to campaign for congressional candidates.
 - (a) You believe that the percentage of votes received by Democratic candidates in each state may be explained by the state's unemployment rate, by whether Al Gore campaigned in that state, and by the state's location in the country in four different regions: Northeast, South, Midwest, and West. Write down a regression model for the determination of the percentage of democratic votes per state that expresses these beliefs. Interpret the coefficient estimates of the model which you would obtain if you would actually use your data to estimate the regression model. Describe how you would test for the following hypotheses using the estimated coefficients and their estimated covariance matrix
 - (a) Gore's campaigning didn't matter.
 - (b) The entire country voted uniformly: there are no regional differences.
 - (c) The Northeast and the Midwest (the "frostbelt") voted uniformly.
 - (d) The frostbelt voted uniformly, the "sunbelt" (the South and the West) voted uniformly, but the frostbelt and the sunbelt did not necessarily vote uniformly.For each one of the hypotheses above, write down the restricted and the unrestricted model and describe how you would use the residuals for each model to test the relevant hypothesis.
 - (b) In addition you may think that the effect of Gore's campaign in a state may be different depending on whether the state is in the Northeast, South, Midwest, or West. Write down a regression model for the determination of the percentage of democratic votes per state that also captures this effect. Interpret the coefficient estimates of the model which you would obtain if you would actually use your data to estimate the regression model. Describe how you would test for the following hypothesis:
 - (a) Gore's appearance had the same effect in all regions.Give three formulas for the test statistic you would use to test the hypothesis above.

3. Suppose that you want to study the relationship between gasoline expenditure and disposable income. You are given quarterly data on real gasoline expenditure (G) and real disposable income (Y), both expressed in 1987 billions of dollars, from 1959:1 to 1992:1. The data are the 2nd and 3rd columns of the data set GAS.TXT, which may be found in the class web site.
 - (a) Plot G and Y against time in the same graph. Define two new series as the natural logarithm of G and Y, say LG and LY. Plot LG and LY against time in the same graph. Comparing the two graphs, what is the effect of taking natural logarithm of the two series?

(b) What are the quarterly rates of growth for G and Y? Are they statistically significant at the 1% and the 5% significance level?

(c) Estimate the model

$$LG_t = \beta_1 + \beta_2 LY_t + \varepsilon_t$$

What is the implied elasticity of demand?

(d) Estimate the model

$$LG_t = \beta_1 + \beta_2 LY_t + \beta_3 T_t + \varepsilon_t$$

What is the implied elasticity of demand?

Economics 143: Problem Set 4

1. In this problem you will study the time-series variation in consumption expenditures in the United States over the period from 1977 (1st quarter) to 1988 (1st quarter). A standard macroeconomic model explains consumption as a general function of disposable income and previous consumption values. However, some opposing points of view are as follows
 1. Consumption is an interstellar process, of which economists know very little, and so it is better to explain consumption by an interstellar activity proxy such as UFO sightings.
 2. Changes in consumption are an interstellar activity, so that consumption should be explained by previous consumption values as well as UFO sightings.
 3. When people decide how much to spend, they consider only their current income, and so consumption should be explained by disposable income only.
 4. People revise their consumption plans via a random process, so that consumption should equal past consumption plus a disturbance.
 5. Consumption is a function of both past consumption and income, so that a dollar increase in income has the same effect as a dollar increase in past consumption.

Use the following regression equation to investigate these various views:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \varepsilon_t$$

Where

$$\begin{aligned} Y_t &= \text{U.S. consumption in time } t \\ X_{2t} &= \text{U.S. disposable personal income in time } t \\ X_{3t} &= \text{U.S. consumption in } t - 1 \\ X_{4t} &= \text{U.S. UFO sightings in time } t \\ \varepsilon_t &= \text{error term} \end{aligned}$$

- (a) Estimate the above equation, using the data from Example 4.3 and your own estimate of UFO sightings.
- (b) Test the following hypotheses by using the appropriate t tests.
 - i. $\beta_4 = 0$ (traditional model)
 - ii. $\beta_2 = 0$ (2)
- (c) Test the following hypotheses by using the appropriate F tests
 - i. $\beta_2 = \beta_3 = 0$ (View 1)
 - ii. $\beta_3 = \beta_4 = 0$ (View 3)
 - iii. $\beta_3 = 1, \beta_2 = \beta_4 = 0$ (View 4)
 - iv. $\beta_4 = 0, \beta_2 = \beta_3$ (View 5)
 - v. $\beta_2 = \beta_3 = \beta_4 = 0$ (No explanatory power from any of the X 's)

Indicate the restricted model appropriate for each of these hypotheses.

2. A four-variable regression using quarterly data from 1958 to 1976 inclusive gave an estimated equation:

$$\hat{Y} = 2.20 + 0.104X_2 - 3.48X_3 + 0.34X_4$$

The explained sum of squares was 109.6 and the error sum of squares 18.48. When the equation was re-estimated with three seasonal dummies added to the specification, the explained sum of squares rose to 114.8.

- (a) Test for the presence of seasonality.

- (b) Two further regressions based on the original specification were run for the subperiods 1958-I to 1968-IV and 1969-I to 1976-IV, yielding error sum of squares of 9.32 and 7.46, respectively. Test for the constancy of the relationship over the two subperiods.
3. This problem uses the data set CPS85 that may be downloaded from the class web site, along with the file ReadCPS that describes the data. The data set is a random sample from the May 1985 Current Population Survey conducted by the U.S. Census Bureau. It contains observations on 12 variables for 534 individuals. (The last 7 variables described in ReadCPS have been dropped.) The first variable in the data set is years of schooling (*EDU*), and the next six entries are 0-1 dummy variables taking on the value 1 if the individual resides in the south (*SOUTH*), is non-white and non-hispanic (*NONWH*), is hispanic (*HISP*), is female (*FE*), is married (*MAR*) and is female and married (*MARRFE*). The next two variables measure potential years of experience (*EX*), computed as age minus years of schooling minus 6, and this potential experience measure squared (*EXSQ*). The next entry is a dummy variable taking on the value 1 if the individual works at a union job (*UNIO*). The next column is the natural logarithm of the individuals average hourly in dollars earnings (*LNWAGE*). The next variable is the individual's age in years (*AGE*). The rest of the variables in the data set will not be used in this problem.

Whenever necessary in the questions below, assume that the assumptions of the Classical Normal Regression model hold.

- (a) Compute the average hourly wage for the entire sample. There are two ways of doing this. Either compute the arithmetic mean of *LNWAGE* and exponentiate it (which gives the geometric mean of average hourly wage, $WAGE = e^{LNWAGE}$), or exponentiate *LNWAGE* for each individual and then compute the arithmetic mean. Are these two measures identical?
- (b) Compute the sample means of the following dummy variables: *SOUTH*, *FE*, *UNIO*, *NONWH*, *HISP*. How many south residents, females, union workers, non-whites & hispanics are in the sample?
- (c) Compute the means and standard deviations of *LNWAGE*, *EDU*, and *EX* for the entire sample, and then by gender (male/female), by race (white/non-white), and by union status (union/non-union). Within each of the three groups sorted by gender, race, and union status, find which subgroup has the highest average *LNWAGE* and the highest dispersion as measured by the standard deviation. Do the same for *EDU*.
- (d) Using Least Squares, estimate the parameters in a simple model where *LNWAGE* is regressed on a constant, years of schooling (*EDU*), and experience (*EX*), and report these along with their estimated standard errors. What does the slope coefficients on *EDU* and *EX* measure? Construct and interpret 95% confidence intervals for these coefficients. Compute and interpret the R^2 coefficient for this model.
- (e) Redo part (d) allowing for gender effects.

Economics 143: Problem Set 5

1. You are estimating a cross-section regression for a sample of 100 cities in the United States in which you hope to explain expenditures on education as a function of the median income in the community, the number of school-age children, and the level of state and federal grants received for educational purposes. Would you expect heteroscedasticity to be a problem in this case? If so, would you use the Goldfeld-Quandt test?
2. You are estimating the relationship between a firm's sales and advertising expenditures in an industry. It becomes apparent to you that half the firms in the industry are large relative to the other half, and you are concerned about the proper estimation technique in such a situation. Assume that the error variances associated with the large firms are twice the error variances associated with the small firms.
 - (a) If you used ordinary least squares to estimate the regression of sales on advertising (assuming that advertising is an independent variable, uncorrelated with the error term), would your estimated parameters be unbiased? Consistent? Efficient?
 - (b) How might you revise the estimation procedure to eliminate or resolve your difficulties?
 - (c) Can you test whether the original error-variance assumption is valid?

3. Five sample observations are

X	4	1	5	8	2
Y	6	3	12	15	4

Assume a linear model, $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, with heteroskedasticity of the form $Var(Y_i) \equiv Var(\varepsilon_i) \equiv \sigma_i^2 = \sigma^2 X_i^2$ where σ^2 is a positive constant. Calculate the OLS and GLS estimates of β_1 and β_2 and the corresponding standard errors.

4. Using CPS85 run the following wage regression and report the estimated standard errors of the coefficients.

$$\begin{aligned}
 LNWAGE = & \alpha_0 + \alpha_F FE + \alpha_U UNIO + \alpha_N NONWH + \alpha_H HISP \\
 & + \gamma_1 EDU + \gamma_2 EX + \gamma_3 EXSQ + \varepsilon
 \end{aligned}
 \tag{3}$$

- (a) Interpret the estimated coefficients. Are all the coefficients statistically significant at the 5% significance level? (As usual, assume that (3) satisfies the assumptions of the Classical Normal Regression model.)
- (b) Perform an F -test of the joint hypothesis that $\alpha_N = \alpha_U = 0$ (i.e. that holding gender, union status, education and experience constant, race doesn't matter for determining wages). Do the test at the 10% significance level.
- (c) An alternative way of performing the test of part (b) is to run the restricted regression:

$$LNWAGE = \alpha_0 + \alpha_F FE + \alpha_U UNIO + \gamma_1 EDU + \gamma_2 EX + \gamma_3 EXSQ + \varepsilon
 \tag{4}$$

and then form the F -statistic using the sum of squared residuals from the unrestricted regression (3) and the restricted regression (4). Perform the test in this manner and compare the value of the F -statistic earlier. Are they the same?

- (d) In performing the tests for race discrimination in wages in parts (b) and (c) above we implicitly assumed that the slope coefficients of the other variables are the same for all races. That is, we assumed that the regression model for each one of the three races, Nonwhites, Hispanics, and the missing reference race, say Whites, have different intercepts but the same slopes for FE , $UNIO$, EDU , EX , $EXSQ$. How would you test whether this assumption is correct, i.e. how would you test whether both intercepts and slopes are different for the three races? EXTRA CREDIT: Perform the test of this last hypothesis at the 10% significance level.

5. Recall the Capital Asset Pricing Model (CAPM), which can be tested by running the OLS regression:

$$r_{at} = \alpha + \beta r_{mt} + \varepsilon_t$$

where $r_a \equiv R_a - R_f$ and $r_m \equiv R_m - R_f$ denote the excess returns on the risky asset and the market portfolio, respectively. Here, R_a is the return of the risky asset, R_f is the return of the risk-free asset, and R_m is the return of the market portfolio. Earlier, we dealt with three files that contain monthly returns from January 1978 till December 1987 ($n = 120$ observations in total) for a risky asset, “Ibm” (IBM stock), a risk-free asset, “Rkfree” (30-day US Treasury Bill), and a market portfolio, “Market”.

- (a) Construct the series $r_a \equiv R_a - R_f$ and $r_m \equiv R_m - R_f$. Run a regression

$$r_{at} = \alpha + \beta r_{mt} + \varepsilon_t$$

Perform a Durbin-Watson test for detecting whether the error process ε_t follows an AR(1).

- (b) Do FGLS using the Cochrane-Orcutt method in order to obtain more efficient estimates of the parameters, assuming that ε_t follows an AR(1), i.e. that $\varepsilon_t = \rho\varepsilon_{t-1} + u_t$, where ρ and u_t satisfy the assumptions discussed in class. (What are these assumptions?)

Economics 143: Problem Set 6

1. Consider the model

$$\begin{aligned}C_t &= \alpha_1 + \alpha_2 Y_t + \varepsilon_t \\I_t &= \beta_1 + \beta_2 Y_t + \beta_3 G_{t-1} + u_t \\Y_t &= C_t + I_t + G_t\end{aligned}$$

- Construct the reduced form system of the model. From the reduced form determine the response of C in the first two periods to a one-unit change in G .
- Is the consumption-function equation identified? Is it overidentified?
- Is the investment equation identified? Overidentified?
- What would happen to your estimated marginal propensity to consume if it has been estimated by using OLS on an equation of the form $C_t = a + bY_t + \varepsilon_t$?

2. Consider the supply-demand model

$$\begin{aligned}Q_t^S &= \alpha_1 + \alpha_2 P_t + \varepsilon_t \\Q_t^D &= \beta_1 + \beta_2 P_t + \beta_3 Y_t + \beta_4 P_{t-1} + u_t \\Q_t^D &= Q_t^S\end{aligned}$$

where $E(\varepsilon_i \varepsilon_j) = 0, i \neq j$, and $E(u_i u_j) = 0, i \neq j$.

- Is the supply equation identified?
- Is the demand equation identified?
- If you were told to estimate the supply equation using instrumental variables, what would you do? Be explicit.
- If you were told to estimate the supply equation using *2SLS*, what would you do? How does this relate to part (c)?

3. Consider the three-equation model system

$$\begin{aligned}Y_1 &= \alpha_1 + \alpha_2 Y_2 + \alpha_4 X_1 + \alpha_5 X_2 + u_1 \\Y_2 &= \beta_1 + \beta_3 Y_3 + \beta_5 X_2 + u_2 \\Y_3 &= \gamma_1 + \gamma_2 Y_2 + u_3\end{aligned}$$

Which of the above equations (if any) are unidentified? Exactly identified? Overidentified?

4. The “Mroz.txt” data file is taken from the 1976 Panel Study of Income Dynamics, and is based on the data for the previous year 1975. Of the 753 observations, the first 428 are for women with positive hours worked in 1975, while the remaining 325 observations are for women who did not work for pay in 1975. The data set consists of 753 observations on 19 variables: THE DATA SET CONSISTS OF 753 OBSERVATIONS ON 19 VARIABLES:

LFP A dummy variable = 1 if woman worked in 1975, else 0

WHR5 Wife’s hours of work in 1975

KL6 Number of children less than 6 years old in household

K618 Number of children between ages 6 and 18 in household

WA Wife’s age

WE Wife’s educational attainment, in years

WW Wife’s average hourly earnings, in 1975 dollars

RPWG Wife's wage reported at the time of the 1976 interview (not the same as the 1975 estimated wage).¹
 HHRS Husband's hours worked in 1975
 HA Husband's age
 HE Husband's educational attainment, in years
 HW Husband's wage, in 1975 dollars
 FAMINC Family income, in 1975 dollars. This variable is used to construct the property income variable.
 MTR This is the marginal tax rate facing the wife, and is taken from published federal tax tables (state and local income taxes are excluded).²
 WMED Wife's mother's educational attainment, in years
 WFED Wife's father's educational attainment, in years
 UN Unemployment rate in county of residence, in percentage points. This taken from bracketed ranges.
 CIT Dummy variable = 1 if live in large city (SMSA), else 0
 AX Actual years of wife's previous labor market experience

- (a) In the model of labor supply estimated by Mroz³, it is assumed that in making her labor supply decisions the wife takes as given the household's entire nonlabor income plus her husband's labor income. Mroz call this sum the wife's property income and computes it as total family income minus the labor income earned by wife. For the entire sample of 753 observations, compute this property income variable PRIN as

$$\text{PRIN} = \text{FAMINC} - \text{WHRS} \cdot \text{WW}.$$

- (b) Restricting your sample to workers, take the natural logarithm of the wife's wage rate variable WW and call this log-transformed variable LWW. Compute the mean and standard deviation of LWW for this sample.
- (c) For the entire sample of 753 observations, construct

$$\begin{aligned} \text{AX2} &= \text{AX} \cdot \text{AX} \\ \text{WA2} &= \text{WA} \cdot \text{WA} \end{aligned}$$

Next, using only the 428 observations from the working sample, regress LWW on a constant term, WA, WE, CIT, AX and AX2. Then, use the parameter estimates from this equation and values of the WA, WE, CIT, AX, and AX2 variable for the 325 women in the nonworking sample to generate the predicted log wage for nonworkers. Call this variable FLWW. Finally, for the entire sample of 753 observations, generate a variable called LWW1 such that LWW1=LWW for the working sample and LWW1=FLWW for the nonworking sample.

- (d) Using OLS estimation procedures, estimate parameters of a linear probability model in which LFP is related linearly to an intercept term, LWW1, KL6, K618, WA, WE, UN, CIT, PRIN, and a stochastic error term. Do signs of these OLS estimated parameters make sense? Next retrieve fitted values from this estimated linear probability model. For how many observations are the fitted values negative? For how many observations are they greater than 1?
- (e) One possible estimation procedure when the dependent variable is binary is logit. With LFP as the dependent variable and with a constant term, LWW1, KL6, K618, WA, WE, UN, CIT, PRIN as explanatory variables, estimate parameters based on a logit maximum likelihood procedure.
- (f) Another possible estimation procedure when the dependent variable is binary is probit. With LFP as the dependent variable and with a constant term, LWW1, KL6, K618, WA, WE, UN, CIT, PRIN as explanatory variables, estimate parameters based on a probit maximum likelihood procedure.

¹To use the subsample with this wage, one needs to select 1975 workers with LFP=1, then select only those women with non-zero RPWG. Only 325 women work in 1975 and have a non-zero RPWG in 1976.

²The taxable income on which this tax rate is calculated includes Social Security, if applicable to wife.

³Mroz (1987, "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions", *Econometrica* 55, 765-799.)